(19) KOREAN INTELLECTUAL PROPERTY OFFICE

KOREAN PATENT ABSTRACTS

(11)Publication number: 1020030035261 A
(43)Date of publication of application: 09.05.2003

(21)Application number: 1020010067244
(22)Date of filing: 30.10.2001

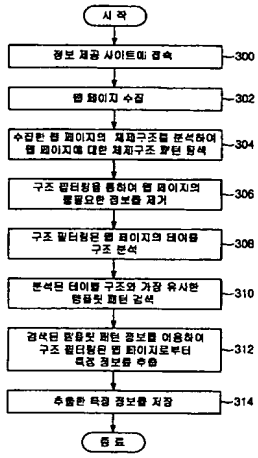
(71)Applicant: HWANG, CHANG HO
PARK, NAM GYU
SONG, HAN BUM

(72)Inventor: HWANG, CHANG HO
PARK, NAM GYU
SONG, HAN BUM

(51)Int. Cl G06F 17/00

(54) METHOD FOR SELECTIVELY EXTRACTING WEB PAGE INFORMATION USING STRUCTURE ANALYSIS

(57) Abstract:



PURPOSE: A method for selectively extracting the web page information using structure analysis is provided to extract the specific information selectively by analyzing a structure of a web page provided from an information providing web site.

CONSTITUTION: After collecting the web page from the information providing web site and searching a layout structure pattern of the collected web page, the unnecessary information is eliminated by performing the structure filtering to the web page as using the information for the layout structure pattern(306). A table structure of the filtered web page is analyzed and a template pattern having the most similar structure with the analyzed table structure is searched from the stored template patterns(310). The specific information is extracted from the filtered page by using the information of the searched template pattern(312).

Legal Status

(19)대한민국특허청(KR)
(12) 공개특허공보(A)

(51) 。Int. Cl. 7
G06F 17/00

(11) 공개번호 특2003-0035261
(43) 공개일자 2003년05월09일

(21) 출원번호 10-2001-0067244
(22) 출원일자 2001년10월30일

(71) 출원인 송한범
부산광역시 동래구 안락1동 429-37

황창호
부산광역시 해운대구 재송2동 시영아파트 6-409

박남규
부산 동래구 온천1동 397-6 번지삼익아파트 2동 1217호

(72) 발명자 송한범
부산광역시동래구안락1동429-37

황창호
부산광역시해운대구재송2동시영아파트6-409

박남규
부산광역시동래구온천2동삼익아파트2-1217

(74) 대리인 특허법인 신성

심사청구 : 없음

(54) 구조분석을 이용한 선택적 웹페이지정보 추출 방법

요약

1. 청구범위에 기재된 발명이 속하는 기술분야

본 발명은 구조분석을 이용한 선택적 웹페이지정보 추출 방법에 관한 것임.

2. 발명이 해결하려고 하는 기술적 과제

본 발명은 정보제공 웹사이트가 제공하는 웹페이지의 구조를 분석하여 특정정보만을 선택적으로 추출할 수 있게 하는, 구조분석을 이용한 선택적 웹페이지정보 추출 방법 및 상기 방법을 실현시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체를 제공하는데 목적이 있음.

3. 발명의 해결 방법의 요지

본 발명은, 정보추출 에이전트(Agent)에 적용되는 웹페이지정보 추출 방법에 있어서, 정보제공 웹사이트로부터 웹페이지를 수집하고, 상기 수집된 웹페이지의 체제(Layout)구조를 분석하여 웹페이지에 대한 체제(Layout)구조패턴을 탐색한 후, 상기 탐색된 체제(Layout)구조패턴에 대한 정보를 이용하여 상기 웹페이지에 구조필터링을 수행하여 불필요한 정보를 제거하는 제 1 단계; 상기 구조필터링된 웹페이지의 테이블(Table)구조를 분석하고, 저장되어 있는 다수의 템플릿 패턴 (Template Patterns) 중에서 상기 분석된 테이블 구조와 가장 유사한 구조를 가지는 템플릿 패턴을 검색하는 제 2 단계; 및 상기 검색된 템플릿 패턴에 대한 정보를 이용하여 상기 제 1 단계에서 필터링된 웹페이지로부터 특정정보를 추출하는 제 3 단계ㄹ르 포함함.

4. 발명의 중요한 용도

본 발명은 웹페이지로부터의 정보추출 등에 이용됨

대표도

도 3

색인어

지능형 정보추출 에이전트, 웹페이지, 체제(Layout)구조, 테이블 구조, 정보추출.

명세서

도면의 간단한 설명

도 1 은 종래의 정보추출 에이전트의 정보추출 방법에 대한 설명도.

도 2 는 본 발명에 따른 지능형 정보추출 에이전트의 일실시예 구성도.

도 3 은 본 발명에 따른 구조분석을 이용한 선택적 웹페이지정보 추출 방법에 대한 일실시예 흐름도.

도 4a 및 도 4b 는 본 발명에 따른 웹페이지의 구조분석 및 구조필터링 방법에 대한 일실시예 설명도.

도 5 는 본 발명에 따른 인터넷 쇼핑몰의 웹페이지로부터 상품정보를 추출하는 방법에 대한 일실시예 설명도.

* 도면의 주요 부분에 대한 부호 설명*

200: 정보제공 웹사이트 202: 인터넷

204: 지능형 정보추출 에이전트 206: 데이터베이스

208: 수집부 210: 분석부

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

본 발명은 웹페이지로부터의 정보추출 방법에 관한 것으로서, 특히 정보제공 웹사이트가 제공하는 웹페이지의 구조를 분석하여 특정정보만을 선택적으로 추출할 수 있게 하는, 구조분석을 이용한 선택적 웹페이지정보 추출 방법 및 상기 방법을 실현시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체에 관한 것이다.

도 1 은 종래의 정보추출 에이전트의 정보추출 방법에 대한 설명도이다.

종래의 정보추출 에이전트는 언어적 분석을 통하여 '주요기사'(100)와 '기획이벤트'(102)사이는 추출대상 정보로 간주하고, 또는 시각적 분석을 통하여 Font size 2와 class=font9udr(104)을 만족하는 텍스트는 추출대상 정보로 간주함으로써 특정정보(106)을 추출한다.

그러나, 종래의 정보추출 에이전트는 텍스트에 기반하는 언어적 구조 및 시각적 구조(예를 들면, 글자, 폰트, 색상, 크

기 등)를 기준으로(즉, HTML소스 상의 특정 단어나 특징을 기준으로) 정보를 분석하기 때문에, 해당 웹페이지의 변경이 이루어져 특정 기준이 없어지면 정확한 정보추출이 어려웠고, 또한 웹사이트의 개별적인 형식으로 인하여 확장성(Wrapper의 생성)에도 많은 제약이 있다는 문제점이 있었다.

발명이 이루고자 하는 기술적 과제

본 발명은, 상기와 같은 문제점을 해결하기 위하여 안출된 것으로, 정보제공 웹사이트가 제공하는 웹페이지의 구조를 분석하여 특정정보만을 선택적으로 추출할 수 있게 하는, 구조분석을 이용한 선택적 웹페이지정보 추출 방법 및 상기 방법을 실현시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체를 제공하는데 그 목적이 있다.

발명의 구성 및 작용

상기의 목적을 달성하기 위한 본 발명은, 정보추출 에이전트(Agent)에 적용되는 웹페이지정보 추출 방법에 있어서, 정보제공 웹사이트로부터 웹페이지를 수집하고, 상기 수집된 웹페이지의 체제(Layout)구조를 분석하여 웹페이지에 대한 체제(Layout)구조패턴을 탐색한 후, 상기 탐색된 체제(Layout)구조패턴에 대한 정보를 이용하여 상기 웹페이지에 구조필터링을 수행하여 불필요한 정보를 제거하는 제 1 단계; 상기 구조필터링된 웹페이지의 테이블(Table)구조를 분석하고, 저장되어 있는 다수의 템플릿 패턴(Template Patterns) 중에서 상기 분석된 테이블 구조와 가장 유사한 구조를 가지는 템플릿 패턴을 검색하는 제 2 단계; 및 상기 검색된 템플릿 패턴에 대한 정보를 이용하여 상기 제 1 단계에서 필터링된 웹페이지로부터 특정정보를 추출하는 제 3 단계를 포함한다.

한편, 본 발명은 웹페이지로부터 정보를 추출하기 위하여, 프로세서를 구비한 정보추출 에이전트(Agent)에, 정보제공 웹사이트로부터 웹페이지를 수집하고, 상기 수집된 웹페이지의 체제(Layout)구조를 분석하여 웹페이지에 대한 체제(Layout)구조패턴을 탐색한 후, 상기 탐색된 체제(Layout)구조패턴에 대한 정보를 이용하여 상기 웹페이지에 구조필터링을 수행하여 불필요한 정보를 제거하는 제 1 기능; 상기 구조필터링된 웹페이지의 테이블(Table)구조를 분석하고, 저장되어 있는 다수의 템플릿 패턴(Template Patterns) 중에서 상기 분석된 테이블 구조와 가장 유사한 구조를 가지는 템플릿 패턴을 검색하는 제 2 기능; 및 상기 검색된 템플릿 패턴에 대한 정보를 이용하여 상기 제 1 기능에서 필터링된 웹페이지로부터 특정정보를 추출하는 제 3 기능을 실현시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체를 제공한다.

이하, 첨부된 도면을 참조하여 본 발명에 따른 바람직한 일실시예를 상세히 설명한다.

도 2 는 본 발명에 따른 지능형 정보추출 에이전트의 일실시예 구성도이다.

지능형 정보추출 에이전트(204)는 정보제공 웹사이트(200)로부터 단순한 웹페이지 전체가 아닌 웹페이지 내부의 특정부분의 정보만을 선택적으로 추출하여 데이터베이스(206)에 저장한다

즉, 지능형 정보추출 에이전트(204)는 정보제공 웹사이트(200)가 제공하는 웹페이지의 구조(예를 들면, 테이블, 셀, 프레임, 위치 등)를 지능화된 추론 능력과 판단력을 통하여 분석하고 선택적으로 정보를 추출한다.

예를 들어, 전자상거래 사이트의 웹페이지로부터 노트북에 대한 정보를 추출하고자 하는 경우, 지능형 정보추출 에이전트(204)는 가격, 노트북 사진, 노트북의 사용후기, 관련 신문기사, 소비자평가 항목 등만을 선택적으로 추출하여 저장한다.

지능형 정보추출 에이전트(204)는 수집부(208)와 분석부(210)로 구성되며, 수집부(208)가 인터넷을 통하여 특정 정보제공처의 웹페이지를 탐색하여 그웹페이지를 수집하면, 분석부(210)는 그 수집한 웹페이지의 구조를 분석하여 정보라고 판단되는 부분을 제외한 나머지 부분을 구조필터링기법을 이용하여 제거함으로써 특정정보를 추출한다.

상기와 같이 구축된 데이터베이스(206)는 정보검색 시스템 등과 연동하여 사용자에게 원하는 정보를 제공하게 된다.

도 3 은 본 발명에 따른 구조분석을 이용한 선택적 웹페이지정보 추출 방법에 대한 일실시예 흐름도이다.

웹브라우저상에 표현되는 웹인터페이스는 단순하게 하나의 웹페이지로 인식되지만, 실제로 지능형 정보추출 에이전트이 대상으로 하는 것은 하이퍼텍스트 생성 언어(HTML: HyperText Markup Language)(이하, HTML라 함)의 소스(Source)이다.

지능형 정보추출 에이전트는 HTML소스의 구조를 파악하고 필요부분만 추출한다. 여기서, 특정정보 추출 과정은 1차 추출과정과 2차 추출과정으로 이루어 지는데, 1차 추출과정에는 체제(Layerout)구조분석과 구조필터링 과정이 해당하고, 2차 추출과정에는 테이블구조분석과 템플릿 패턴 검색 과정이 해당된다.

이하, 신문기사의 리스트를 추출하는 과정을 예로 들어 설명하면, 다음과 같다.

지능형 정보추출 에이전트(204)가 정보제공 사이트(200)에 접속하여(300), 그 정보제공 사이트(200)가 제공하는 웹페이지를 수집한다(302).

지능형 정보추출 에이전트(204)는 수집된 웹페이지의 체제(Layerout)구조를 분석한다(304). 신문사 사이트가 제공하는 웹페이지는 신문기사 리스트뿐만 아니라 상단메뉴와 좌측메뉴, 기타 광고, 그리고 불필요한 정보 등으로 이루어져 있는데, 체제(Layerout)구조 분석 과정에서는 웹페이지 체제(Layerout)구조상 반복되는 패턴과 반복되지 않는 패턴을 비교하여 신문기사 리스트를 정확하게 추출해 낸다.

예를 들어, 상단메뉴나 좌측메뉴, 및 기타 정보들의 구조적인 위치나 형태 등은 일반적으로 반복되기 때문에, 웹페이지 별로 체제(Layerout)구조를 비교하면 반드시 반복되는 패턴이 발생하는데, 이렇게 반복되는 패턴은 불필요한 정보로 간주하게 된다.

지능형 정보추출 에이전트(204)는 웹페이지의 체제(Layerout)구조 분석 과정 (304)을 통하여 획득된 체제(Layerout)구조분석 정보를 이용하여 웹페이지의 구조필터링을 수행한다(306). 여기서, 구조필터링이란 필요한 정보만 남기고, 나머지 불필요한 HTML소스를 삭제하는것이다.

이후, 지능형 정보추출 에이전트(204)는 2차 추출과정을 수행하게 되는데, 1차 추출과정(304, 306)을 통하여 가공된 정보에 대하여 테이블구조 분석 및 유사 템플릿 패턴 검색을 수행한다(308, 310).

1차 추출과정에 의하여 가공된 정보는 일단 '이것이 정보이다'라는 속성만 가지고 있을 뿐, 세부적 사항인 기사제목, 본문내용, 작성일자, 작성자 등과 관련된 사항을 내포하고 있지는 않다. 따라서, 이러한 세부적인 메타데이터 개념의 정 의를 내리는 과정이 테이블구조 분석 및 유사 템플릿 패턴 검색 과정인 것이다.

일반적으로 정보로 간주되는 HTML 소스의 속성은 테이블(td 및 tr이라는 태그로 이루어짐)의 형태를 이루고 있기 때문에, HTML 소스의 속성 분석은 테이블이 어떻게 이루어져 있는가를 분석하는 과정이다(308).

예를 들어, 테이블이 횡으로 3인가 혹은 4인가에 따라 다른 패턴의 모델이 적용되며, 또한 이러한 과정에서도 흔하지 않은 태그(li 등)의 사용 등이 이루어지는 경우도 있어 테이블의 구조를 정확하게 분석하는 것이 중요하다.

상기의 같은 과정을 통하여, 정보 HTML소스의 형태는 패턴을 분석하기 용이하게 변형되며, 그 변형된 정보 HTML 소스는 이미 구축되어져 있는 템플릿 패턴과 비교하여 가장 유사한 템플릿 패턴을 찾는다(310).

요컨대, 테이블구조분석 과정은 HTML소스의 테이블의 구조를 분석하는 것이고, 유사 템플릿 패턴 검색 과정은 이미 데이터베이스에 저장되어 있는 템플릿 패턴 중에서 그 분석된 테이블 구조와 가장 유사한 템플릿 패턴을 찾는다.

1차 추출과정에 의하여 가공된 정보는 '308' 및 '310'을 통하여 추출된 패턴(즉, 테이블 구조와 가장 유사한 템플릿 패턴)을 이용하여 특정정보를 추출하여(312), 데이터베이스에 저장한다(314).

도 4a 및 도 4b 는 본 발명에 따른 웹페이지의 체제(Layerout)구조분석 및 구조필터링 방법에 대한 일실시예 설명도이다.

도 4a 는 웹페이지의 체제구조를 분석하여(즉, 다수의 웹페이지를 비교하여) , 반복되는 패턴을 찾는 과정을 나타내며 , 도 4b 는 체제(Layerout)구조 분석에 의 하여 찾아낸 반복패턴을 불필요한 정보로 취급하여 웹페이지(400)로부터 필터링(제거)함으로써 필요한 정보(402)만을 추출하는 과정을 나타낸다.

도 5 는 본 발명에 따른 인터넷 쇼핑몰의 웹페이지로부터 상품정보를 추출하는 방법에 대한 일실시예 설명도이다.

도면에 도시된 바와 같이, 지능형 정보추출 에이전트(204)가 인터넷 쇼핑몰의 웹페이지로부터 상품명(506), 상품사진(500), 가격(504), 제품특징(502) 등의 상품정보만을 추출하는 과정을 나타낸다.

또한, 지능형 정보추출 에이전트(204)는 개별 웹사이트의 독특한 포맷관행이나 웹페이지 언어(외국 사이트의 경우)에 관계없이 지능적으로 웹페이지의 구조를 분석하여 특정정보를 선택적으로 추출한다.

이상에서 설명한 본 발명은, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 있어 본 발명의 기술적 사상을 벗어나지 않는 범위내에서 여러 가지 치환, 변형 및 변경이 가능하므로 전술한 실시예 및 첨부된 도면에 의해 한정되는 것이 아니다.

발명의 효과

상기와 같은 본 발명은, 정보제공 웹사이트에서 특정정보만을 선택적으로 추출할 수 있게 하는 효과가 있다. 즉, 본 발명은, 기사를 제공하는 각 신문사 웹사이트로부터 정치, 경제, 연예, 사회면 등 특정 카테고리 내의 부분정보만을 추출하거나, 또는 인터넷 쇼핑몰로부터 상품명, 가격, 상품사진, 제품특징 등만을 선택적 으로 추출할 수 있게 하는 효과가 있다.

또한, 본 발명은, 언어적/시각적 특징을 기준으로 하지 않고 HTML소스 자체의 구조를 파악하여 이를 기준으로 정보를 추출하기 때문에, 범용적이고 확장성이 강한 정보추출 에이전트를 개발할 수 있게 하는 효과가 있다.

또한, 본 발명은, 인터넷 상에서 발생할 수 있는 여러 가지 비즈니스와 웹서비스를 가능하게 만드는 지능형 솔루션으로서, 정보검색시스템, 정보통합시스템, 전자상거래, 컨텐츠산업(Syndication), 고객관계관리(eCRM), 기업정보포털(EIP), 개인포털(PIP), P2P그룹웨어 등 다양한 분야에 적용될 수 있는 효과가 있다.

(57) 청구의 범위

청구항 1.
정보추출 에이전트(Agent)에 적용되는 웹페이지정보 추출 방법에 있어서,

정보제공 웹사이트로부터 웹페이지를 수집하고, 상기 수집된 웹페이지의 체제(Layout)구조를 분석하여 웹페이지에 대한 체제(Layout)구조패턴을 탐색한 후, 상기 탐색된 체제(Layout)구조패턴에 대한 정보를 이용하여 상기 웹페이지에 구조필터링을 수행하여 불필요한 정보를 제거하는 제 1 단계;

상기 구조필터링된 웹페이지의 테이블(Table)구조를 분석하고, 저장되어 있는 다수의 템플릿 패턴(Template Patterns) 중에서 상기 분석된 테이블 구조와 가장 유사한 구조를 가지는 템플릿 패턴을 검색하는 제 2 단계; 및

상기 검색된 템플릿 패턴에 대한 정보를 이용하여 상기 제 1 단계에서 필터링된 웹페이지로부터 특정정보를 추출하는 제 3 단계

를 포함하는 구조분석을 이용한 선택적 웹페이지정보 추출 방법.

청구항 2.
제 1 항에 있어서,

상기 제 1 단계의 웹페이지의 체제(Layout)구조분석은,

상기 웹페이지를 작성한 하이퍼텍스트생성언어(HTML)의 소스(Source)의 체제 (Layout)구조를 분석하는 것을 특징으로 하는 구조분석을 이용한 선택적 웹페이지정보 추출 방법.

청구항 3.
제 1 항 또는 제 2 항에 있어서,

상기 제 1 단계의 구조필터링은,

상기 정보제공 웹사이트가 제공하는 다수의 웹페이지마다 반복되는 체제 (Layout)구조패턴을 불필요한 정보로 간주하여 제거하는 것을 특징으로 하는 구조분석을 이용한 선택적 웹페이지정보 추출 방법.

청구항 4.
웹페이지로부터 정보를 추출하기 위하여, 프로세서를 구비한 정보추출 에이전트(Agent)에,

정보제공 웹사이트로부터 웹페이지를 수집하고, 상기 수집된 웹페이지의 체제(Layout)구조를 분석하여 웹페이지에

대한 체제(Layout)구조패턴을 탐색한 후, 상기 탐색된 체제(Layout)구조패턴에 대한 정보를 이용하여 상기 웹페이지에 구조필터링을 수행하여 불필요한 정보를 제거하는 제 1 기능;

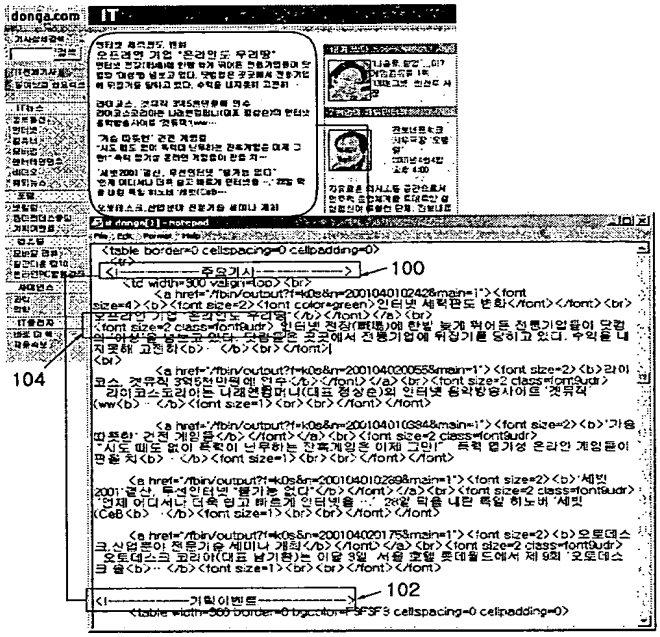
상기 구조필터링된 웹페이지의 테이블(Table)구조를 분석하고, 저장되어 있는 다수의 템플릿 패턴(Template Patterns) 중에서 상기 분석된 테이블 구조와 가장 유사한 구조를 가지는 템플릿 패턴을 검색하는 제 2 기능; 및

상기 검색된 템플릿 패턴에 대한 정보를 이용하여 상기 제 1 기능에서 필터링된 웹페이지로부터 특정정보를 추출하는 제 3 기능

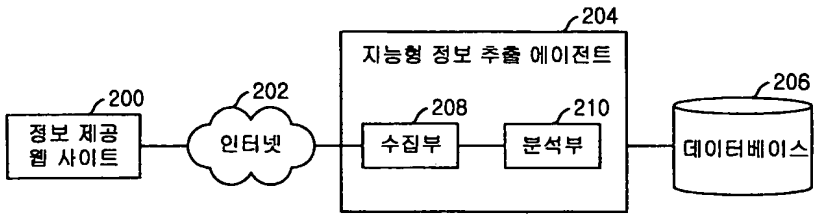
을 실현시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

도면

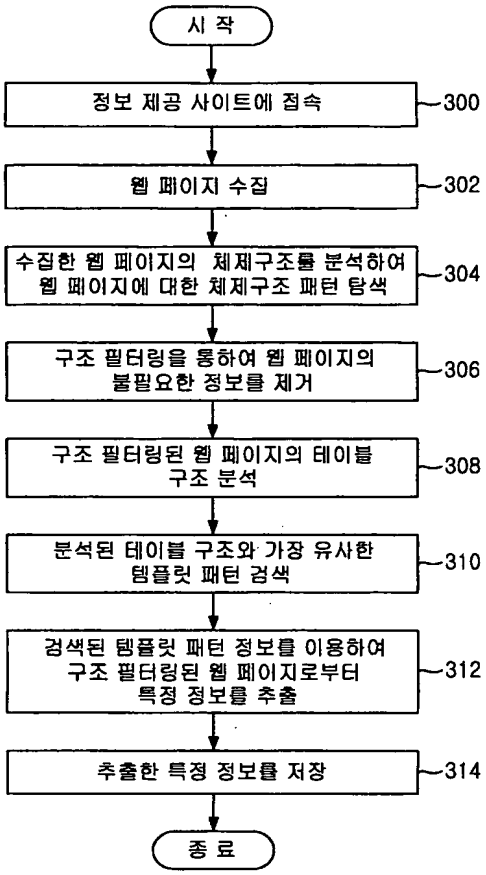
도면1



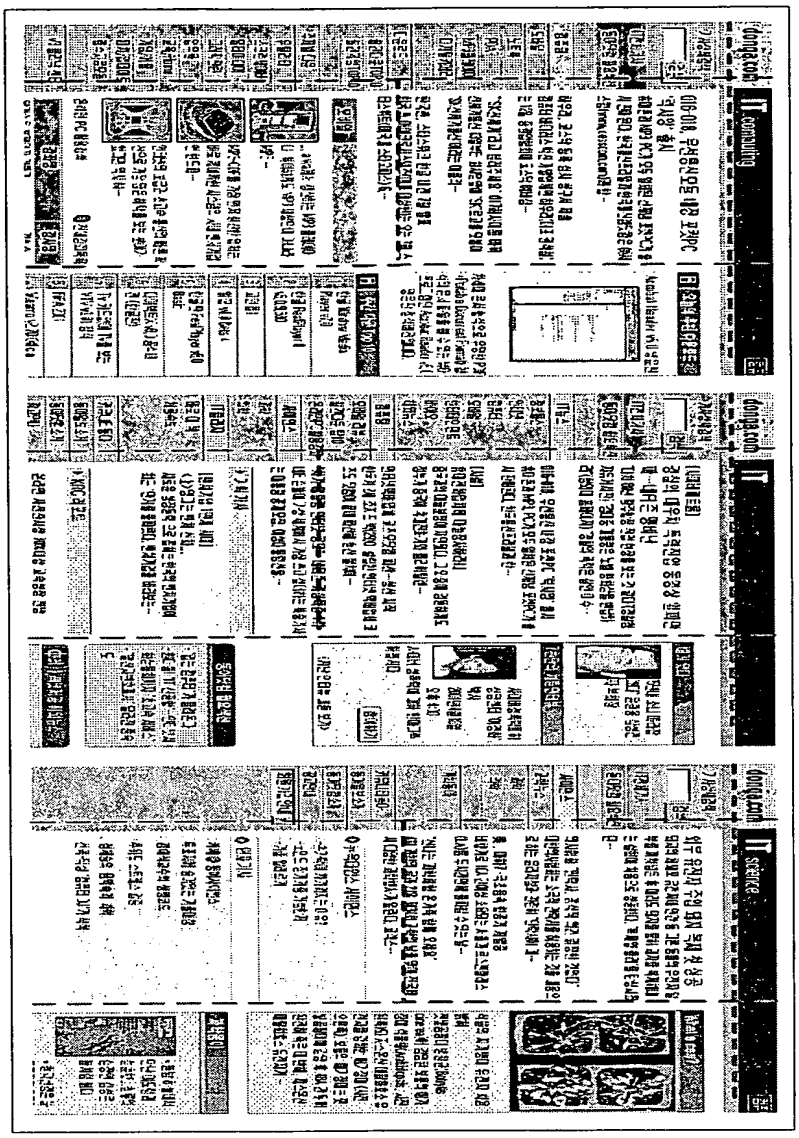
도면2



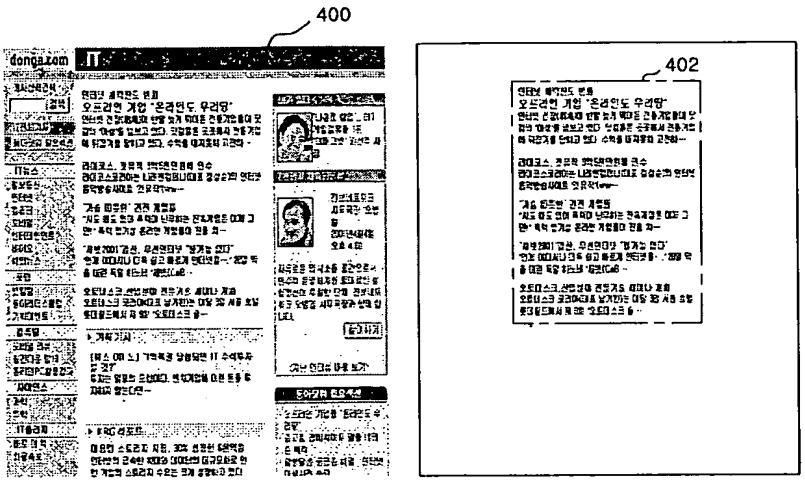
도면3


시 작
정보 제공 사이트에 접속
300
웹 페이지 수집
302
수집한 웹 페이지의 체제구조를 분석하여 웹 페이지에 대한 체제구조 패턴 탐색
304
구조 필터링을 통하여 웹 페이지의 불필요한 정보를 제거
306
구조 필터링된 웹 페이지의 테이블 구조 분석
308
분석된 테이블 구조와 가장 유사한 템플릿 패턴 검색
310
검색된 템플릿 패턴 정보를 이용하여 구조 필터링된 웹 페이지로부터 특정 정보를 추출
312
추출한 특정 정보를 저장
314
종 료

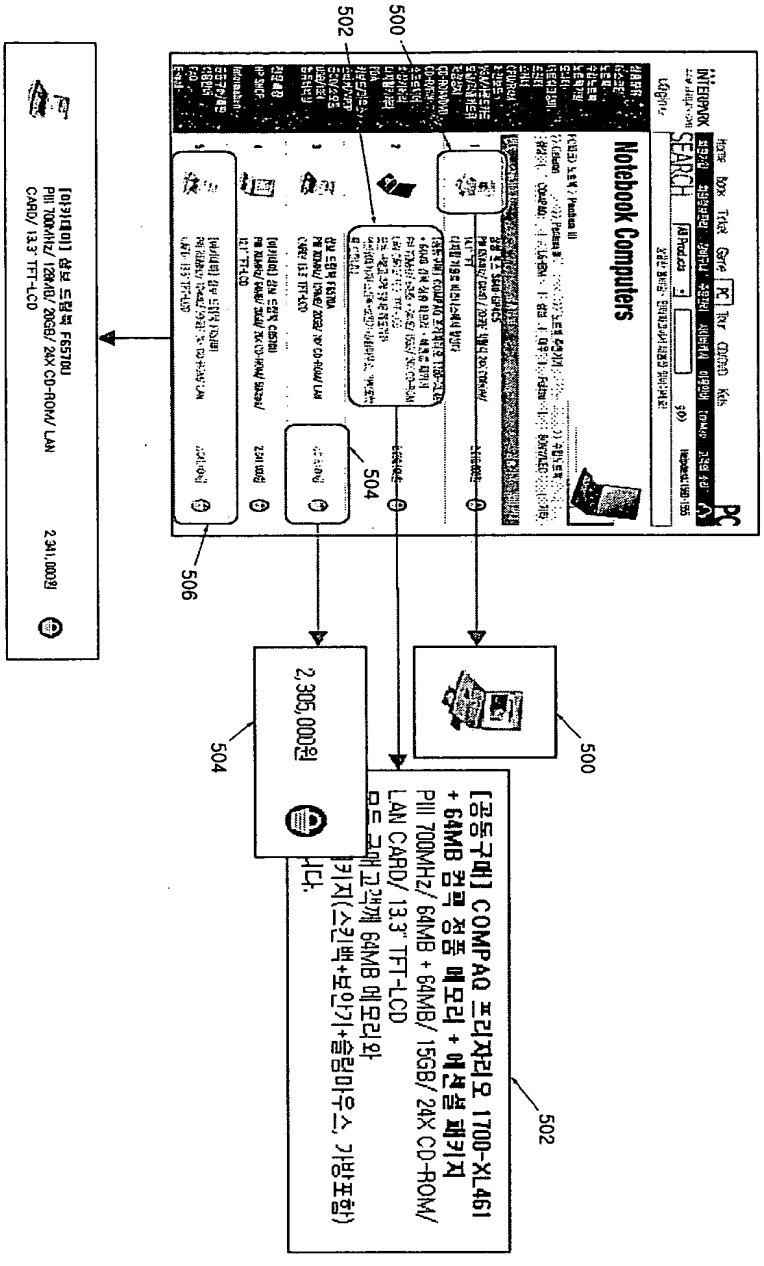
도면4a



도면4b



도면5


500
502
504
506
Notebook Computers
[아이테크] 삼보 드림북 F6570J
PIII 700MHz/ 128MB/ 20GB/ 24X CD-ROM/ LAN CARD/ 13.3" TFT-LCD
2,341,000원
2,305,000원
[공동구매] COMPAQ 프리자리오 1700-XL461 + 64MB 컴팩 정품 메모리 + 여성을 패키지
PIII 700MHz/ 64MB + 64MB/ 15GB/ 24X CD-ROM/ LAN CARD/ 13.3" TFT-LCD
모든 구매고객께 64MB 메모리와
키지(스킨백+보안기+슬림마우스, 가방포함)
다.